Psicometria con Laboratorio di SPSS 2

Regressione lineare semplice

(v. 1.3, 7 novembre 2022)

Germano Rossi[1]
germano.rossi@unimib.it

[1]Dipartimento di Psicologia, Università di Milano-Bicocca

2017-18

Sommario

1. Regressione lineare semplice
2. In SPSS
3. Regressione semplice matriciale (M)
4. Riassunto terminologico

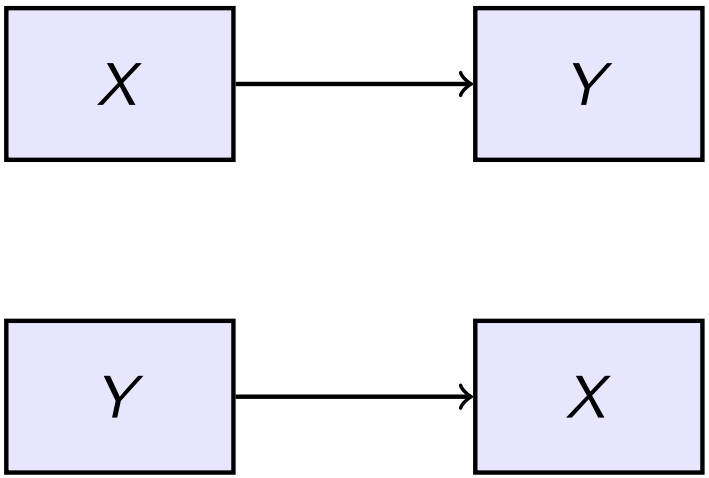
ATTENZIONE: Tutte le slide con una lettera nell'angolo destro del titolo, indicano: A = delle informazioni Avanzate da leggere successivamente dopo aver compreso le altre; M = una parte che utilizza l'algebra Matriciale che è possibile saltare.

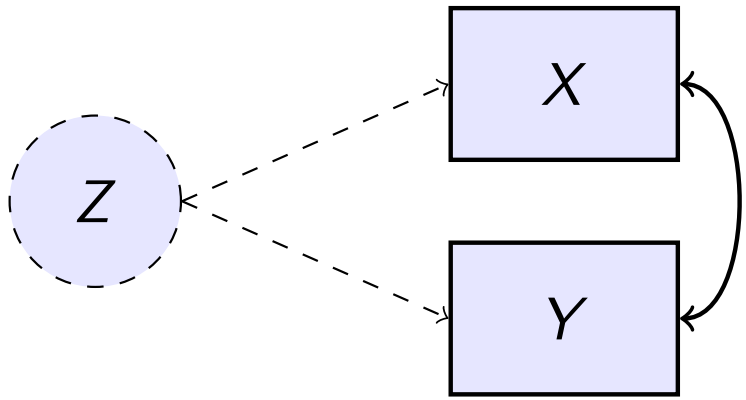
Sommario locale

1. Regressione lineare semplice
   - Differenza fra correlazione e influenza
   - Cos'è la regressione semplice
   - Stime, residui, pendenze
   - Test di significatività
2. In SPSS
3. Regressione semplice matriciale (M)
4. Riassunto terminologico

Correlazione e influenza

- La correlazione fra due variabili (X e Y) implica un'influenza reciproca o "associazione"
- Vi sono diverse possibili spiegazioni
  - X spiega Y [regressione semplice]
  - Y spiega X [regressione semplice]
  - Y è spiegata da X_1, X_2, X_3... X_n [regressione multipla]
  - X e Y sono spiegati da Z (analisi fattoriale)
  - X e Y sono spiegati da Z_1, Z_2, Z_3... Z_n (equazione strutturale)





Significato di Influenza

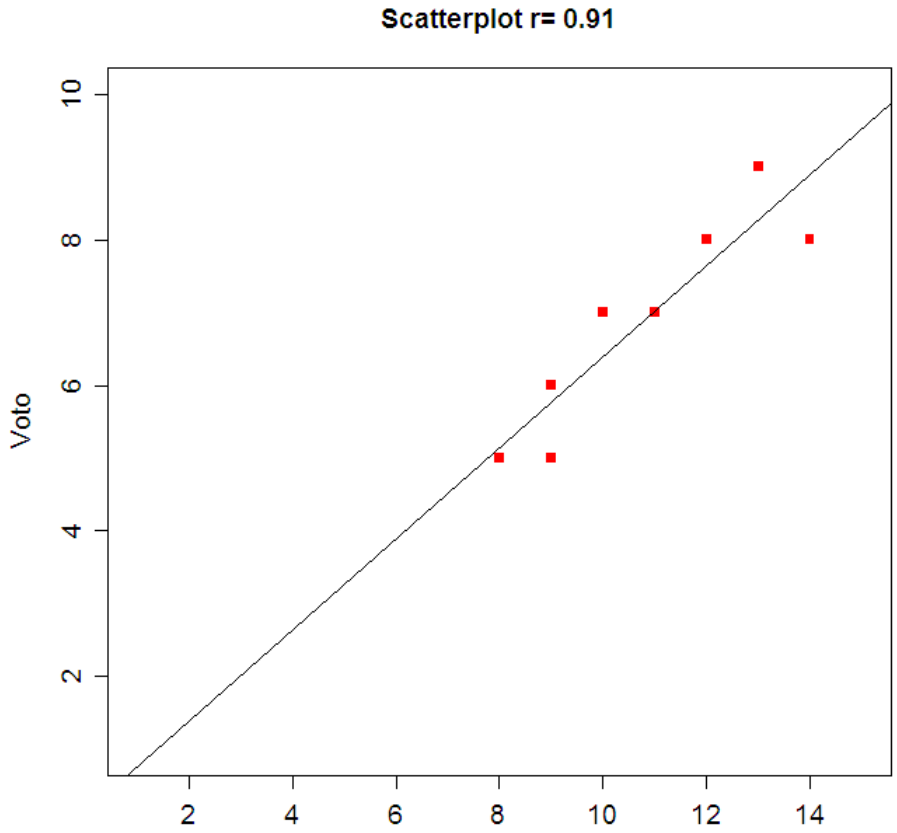
- "Influenza" significa che una variabile **Indipendente** spiega, influenza oppure causa qualche tipo di effetto sulla **Dipendente**
- Spiegazione, influenza e Causa hanno pressapoco lo stesso significato (in ambito statistico) ma si preferiscono i primi due termini
- Ipotizziamo che il **Voto** di un insegnamento sia influenzato dalle **Ore di studio**.
- Se non ci fosse un'influenza, il *Voto* sarebbe uguale al suo valore atteso $E(\text{voto})$, cioè la media
- Un diverso numero di *Ore di studio* possono far aumentare o diminuire il *Voto* o, comunque, possono "far variare" il suo punteggio
- La correlazione ci dice se c'è un legame fra le due variabili
- Il valore della correlazione, indica l'ampiezza del legame

Cos'è la regressione semplice

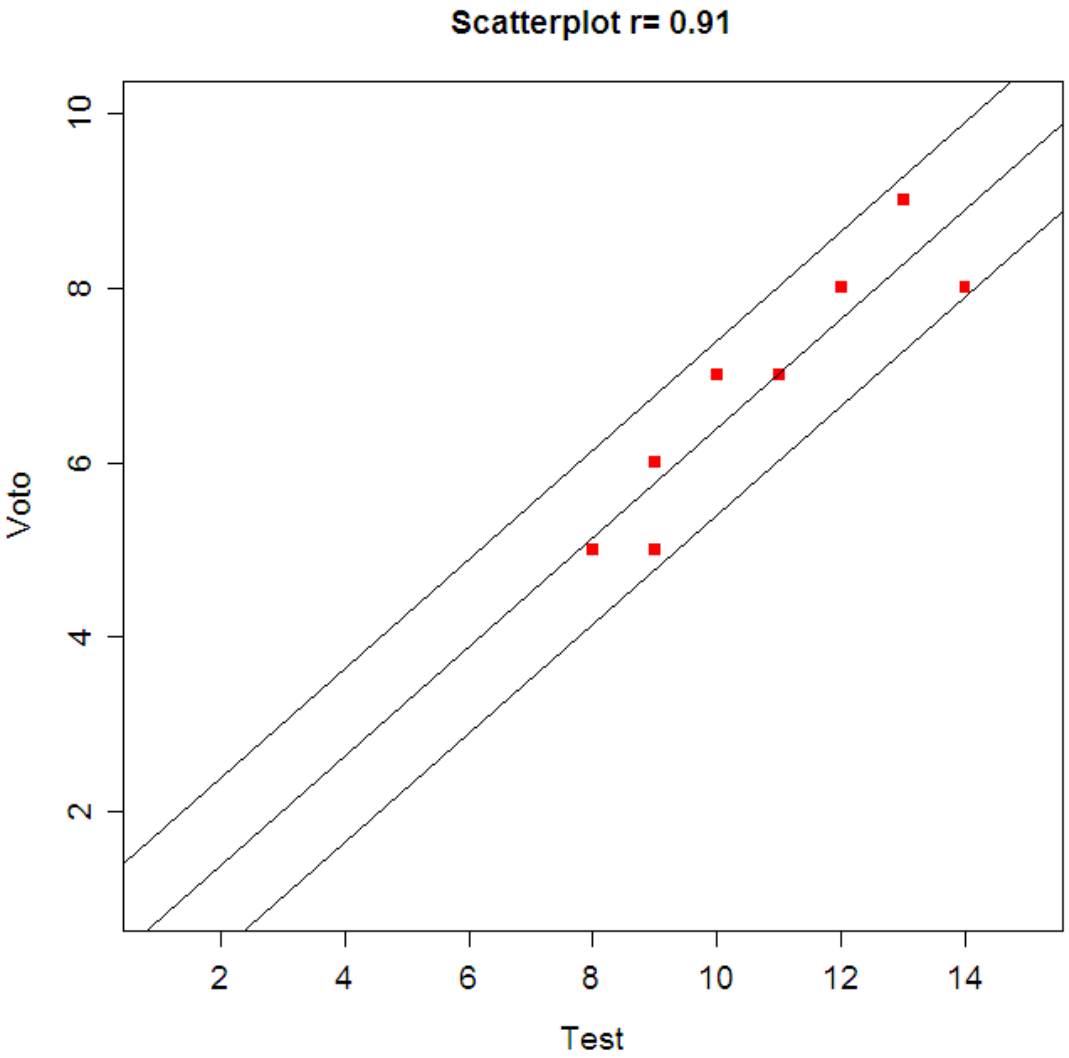
- Con la regressione lineare semplice cerchiamo di vedere se la (cor)relazione fra due variabili (ad es. *Test di metà anno* e *Voto finale*) può essere spiegata tramite l'equazione di una retta
- Quale variabile sia la dipendente e quale l'indipendente, è una scelta teorica (ipotizzo che il test di metà anno spieghi il voto finale)

Usando il file `testVoto.xlsx`.

	Test	Voto
A	12	8
B	10	7
C	14	8
D	9	5
E	9	6
F	13	9
G	11	7
H	8	5

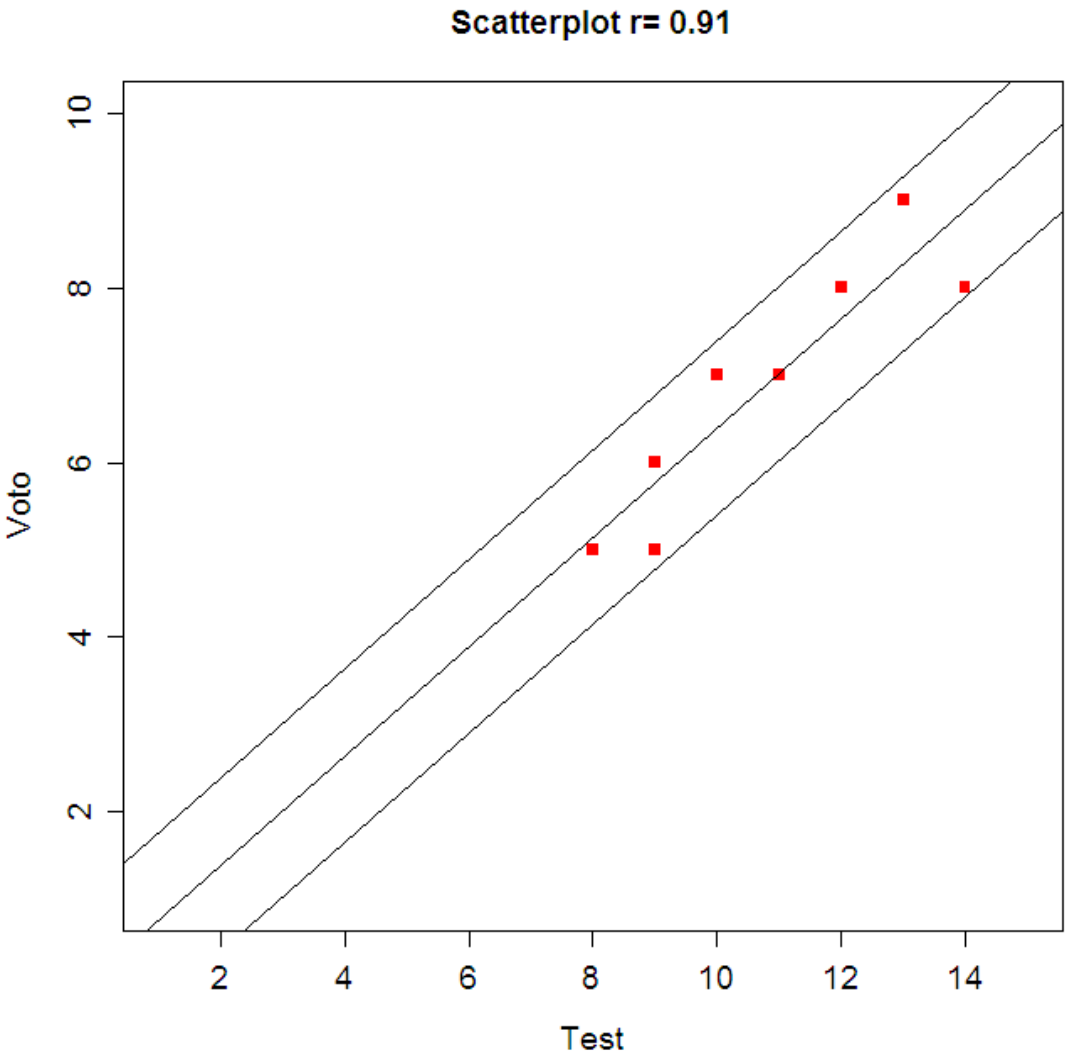


Scatterplot fra Test e Voto



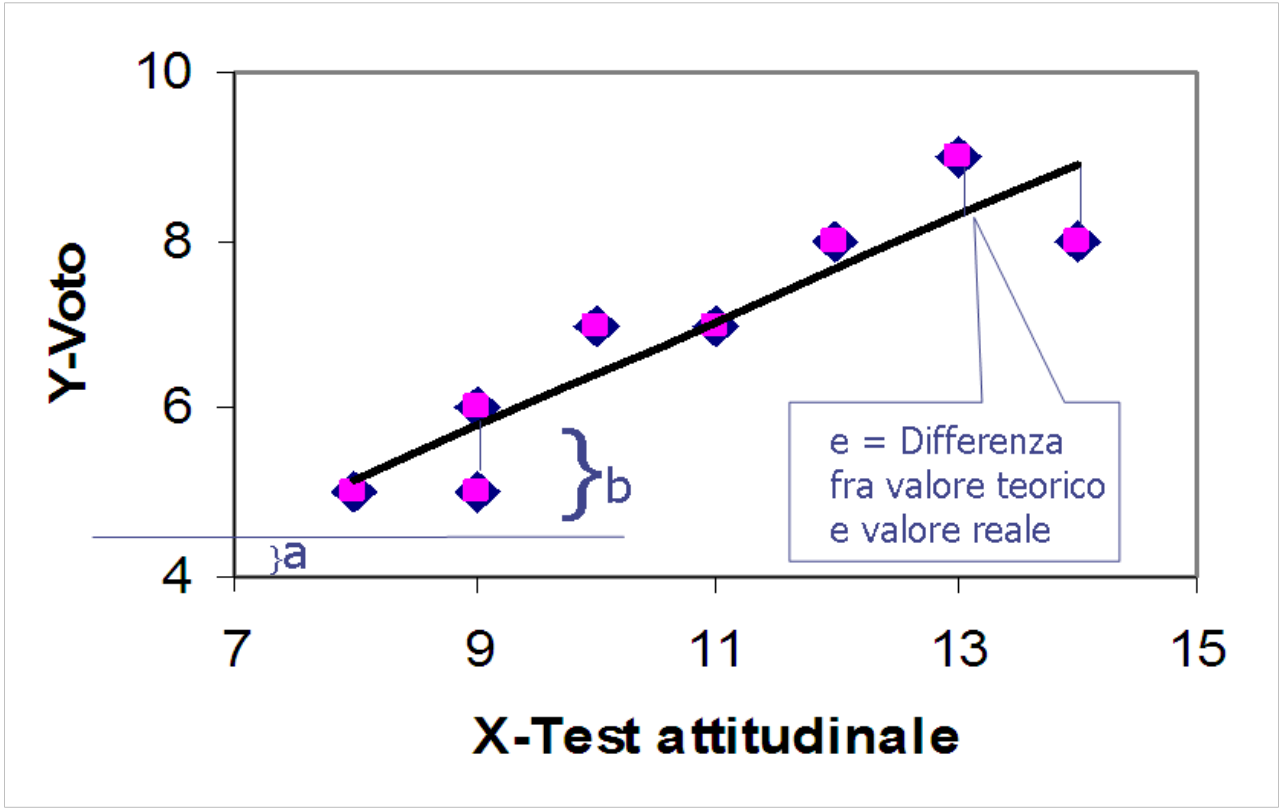
- Cercando una funzione matematica che "approssimi" i punti, posso usare una retta
- ci sono però diverse rette che posso usare
- ciascuna è più vicina a certi punti rispetto alle altre
- qual è la migliore?

Scatterplot fra Test e Voto



- Cercando una funzione matematica che "approssimi" i punti, posso usare una retta
- ci sono però diverse rette che posso usare
- ciascuna è più vicina a certi punti rispetto alle altre
- qual è la migliore?
- quella che, contemporaneamente, è alla minor distanza possibile da tutti i punti osservati

Grafico della retta



- La formula della retta è

$$Y_i = a + bX_i$$

oppure

$$Y_i = b_0 + b_1 X_i$$

- dove X (l'indipendente) e Y (la dipendente) sono le variabili misurate

- b o b_1 è la pendenza della retta
- a o b_0 è l'intercetta sull'asse delle ordinate

Equazione

- In teoria, la formula della retta che interpola meglio è:

$$Y_i = a + bX_i$$

- ma i valori i valori che calcoleremmo sono perfettamente posizionati sulla retta, mentre invece non lo sono affatto
- per cui consideriamo i risultati della retta come una stima (indicati con Y'_i o con $\widehat{Y}_i$) di dove cadrebbero i punti reali se fosse vera l'equazione della retta

$$\widehat{Y}_i = a + bX_i$$

- Per ottenere i veri valori osservati, bisogna aggiungere una *variabile d'errore* che contenga i valori che correggono la retta
- questa è l'equazione esatta, perché considera anche l'errore che permette di aggiustare i dati:

$$Y_i = a + bX_i + e_i$$

Riepilogo rette

$$Y_i = bX_i + a + \varepsilon_i$$

Variabile dipendente, spiegata, valore osservato

inclinazione

variabile indipendente

intercetta

errore

Stima di y, valore predetto

$$\hat{Y}_i = bX_i + a$$

Riepilogo valori

Abbiamo

- Y, è il punto associato all'intersezione fra Test e Voto (pallini blu)
- $\widehat{Y}$, è la stima del punto Y ottenuto tramite la retta (pallini rossi)
- $|Y - \widehat{Y}|$, è la distanza fra il valore reale (Y) e quello stimato ($\widehat{Y}$, righe nere)
- Quest'ultimo, è anche chiamato "errore" (e_i)

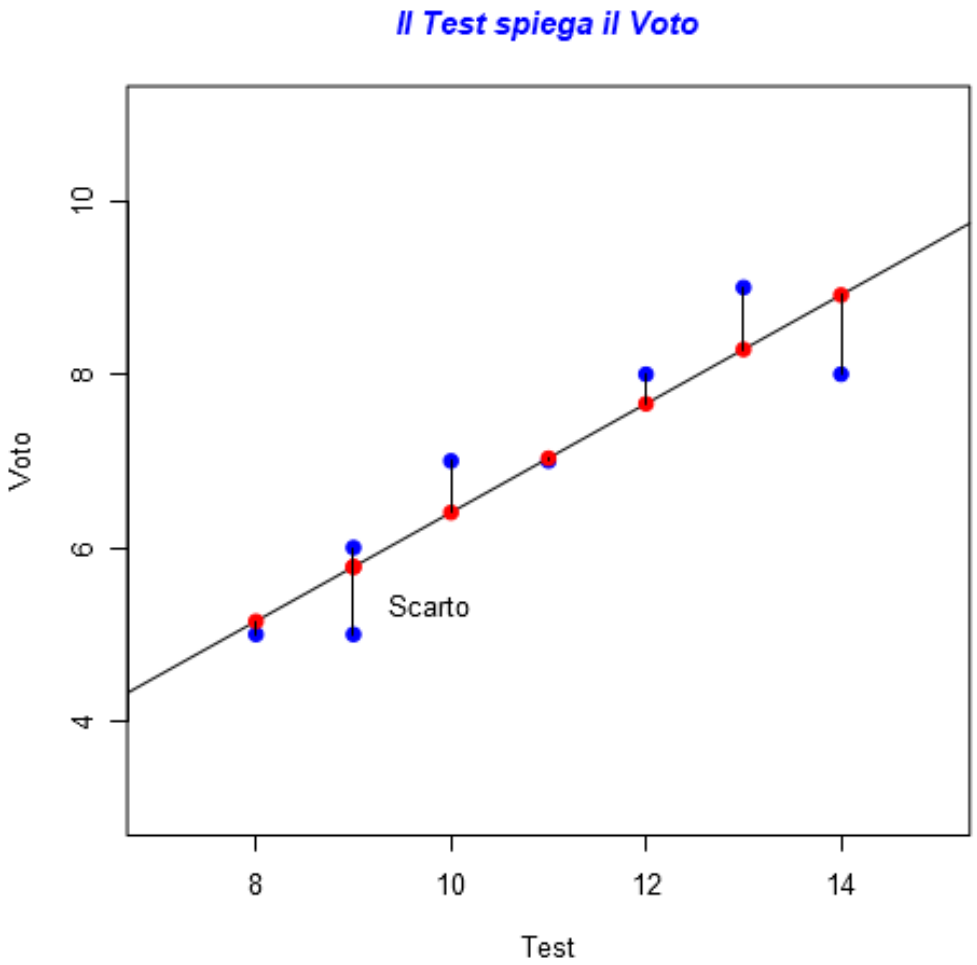
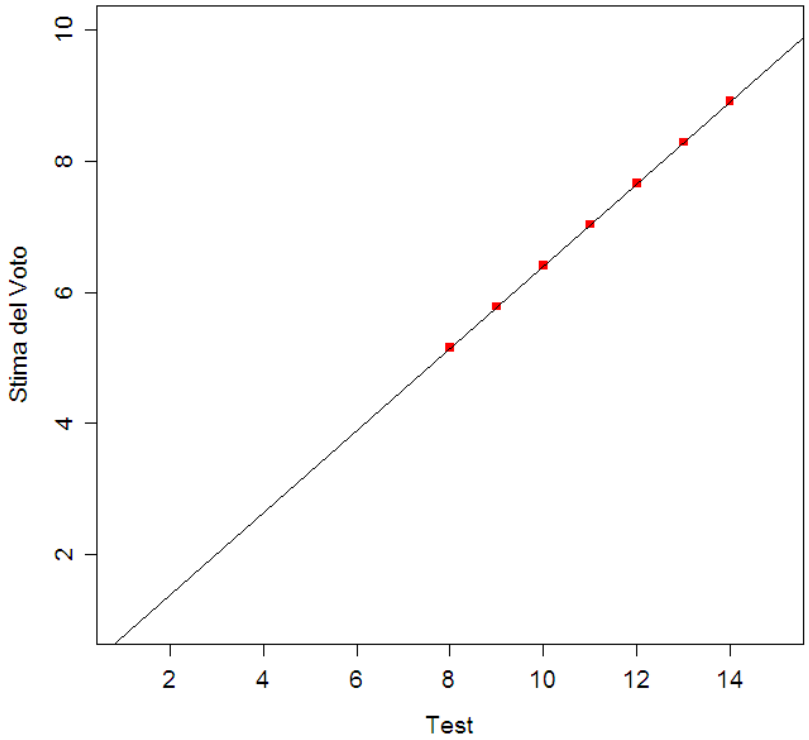


Grafico fra X e Y'



- Se stimo i valori y usando la retta
- i punti cadono esattamente sulla retta
- Fra tutte le possibili rette, selezioniamo la migliore, ovvero quella che è alla minor distanza possibile da tutti i punti osservati
- ovvero, in cui la somma degli errori è la minima possibile
- Obiettivo: minimizzare $\sum e_i$
- $e_i = Y_i - \widehat{Y}_i = Y_i - bX_i + a$

- Bisogna che questi errori siano i più piccoli possibili e quindi usiamo il "metodo dei minimi quadrati" o "minimi residui"

Formule algebriche

Per stimare b ovvero la pendenza, usiamo:

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\text{cov(X,Y)}}{\text{var(X)}} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$

$$= \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2}$$

- in verde: covarianza diviso varianza di X
- in rosa: correlazione moltiplicata per il rapporto fra le deviazioni standard di Y e di X

Per stimare a ovvero l'intercetta, usiamo:

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{N} - b\frac{\sum X}{N}$$

Esempio

	X-Test	Y-Voto	X^2	Y^2	XY
A	12	8	144	64	96
B	10	7	100	49	70
C	14	8	196	64	112
D	9	5	81	25	45
E	9	6	81	36	54
F	13	9	169	81	117
G	11	7	121	49	77
H	8	5	64	25	40
Somma	86	55	956	393	611
Media	10,75	6,875			

$$a = \bar{Y} - b\bar{X}$$

$$a = 6.875 - 0,627$$

$$\times 10.75 = 0.135$$

$$b = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2}$$

$$b = \frac{8 \times 611 - 86 \times 55}{8 \times 956 - (86)^2} = \frac{4888 - 4730}{7648 - 7396} = \frac{158}{252} = 0,627$$

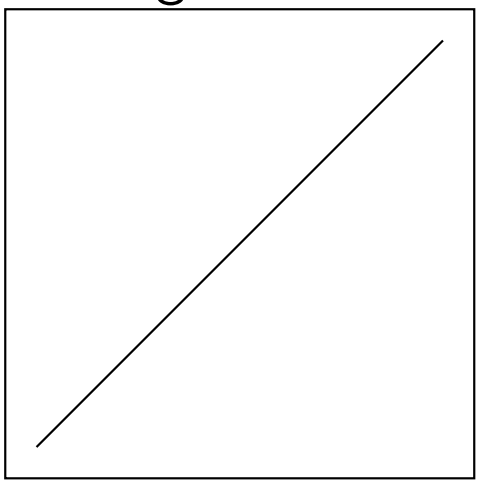
Stime di Y e residui

	Test X	Voto Y	Eq. $a + bX_i$	Stimati $\hat{Y}$	Residui $e = Y - \hat{Y}$
A	12	8	$0.135 + 0.627(12)$	7.659	0.341
B	10	7	$0.135 + 0.627(10)$	6.405	0.595
C	14	8	$0.135 + 0.627(14)$	8.913	-0.913
D	9	5	$0.135 + 0.627(9)$	5.778	-0.778
E	9	6	$0.135 + 0.627(9)$	5.778	0.222
F	13	9	$0.135 + 0.627(13)$	8.286	0.714
G	11	7	$0.135 + 0.627(11)$	7.032	-0.032
H	8	5	$0.135 + 0.627(8)$	5.151	-0.151

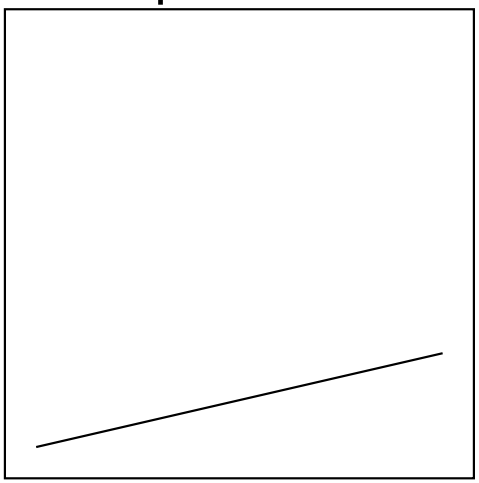
Confronto delle pendenze

- Il coefficiente angolare dipende dal modo in cui è espressa la variabile X
- non si può dire se sia piccolo o grande se non conoscendo la gamma (differenza fra massimo e minimo) di X
- oppure facendo una rappresentazione grafica

grande



piccola



Con dati standardizzati

- Se usiamo X e Y trasformati in punti z, la formula della retta cambia in

$$z_{\hat{y}} = r z_x$$

perché tutti dati sono espressi con media 0 e dev.st 1, quindi

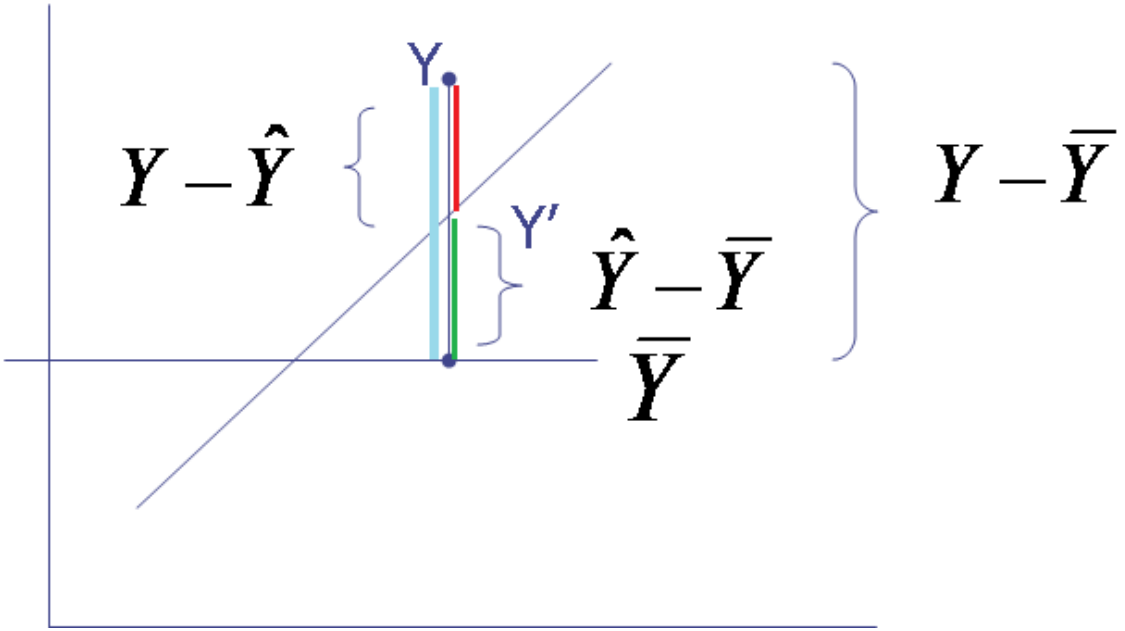
$$\mathrm{b} = r \frac{s_y}{s_x}$$

diventa

$$b^* = r$$

- b^* è la pendenza standardizzata
- l'intercetta è 0 perché $a = \bar{Y} - b\bar{X}$ e le medie sono 0
- Ne consegue che:
- In una regressione lineare semplice
- la pendenza standardizzata coincide con la correlazione fra le due variabili

I residui e le loro sommatorie



- Senza interferenze esterne, $Y_i = E(Y) = \bar{Y}$
- $Y - \bar{Y}$ può essere diviso in due parti
- L'introduzione di X, giustifica la parte $\hat{Y} - \bar{Y}$

- $Y - \hat{Y}$ non abbiamo idea di cosa lo produca
- Possiamo dire che $Y - \bar{Y}$ è divisibile in una parte spiegata da X $\hat{Y} - \bar{Y}$ e in una parte non spiegata (il residuo) $Y - \hat{Y}$

Residui e varianze

- Dal momento che la somma al quadrato degli scarti dalla media corrisponde alla varianza... possiamo trasformare la relazione

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$$

in

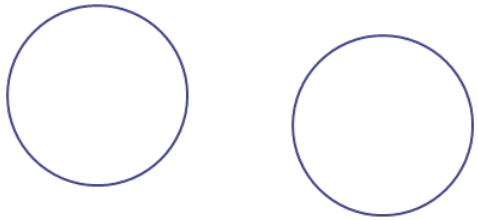
$$\underbrace{\sum(Y - \bar{Y})^2}_{\text{totale}} = \underbrace{\sum(Y - \hat{Y})^2}_{\text{non spiegata}} + \underbrace{\sum(\hat{Y} - \bar{Y})^2}_{\text{spiegata}}$$

- facendo il rapporto fra la varianza spiegata e quella totale, la possiamo esprimere come proporzione di varianza spiegata

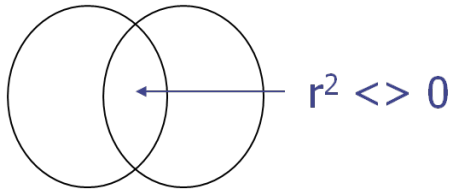
$$r^2 = (r)^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{\sum(Y - \bar{Y})^2 - \sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

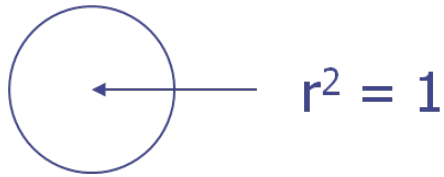
Proporzione di varianza spiegata

- La proporzione di varianza spiegata è anche chiamata "Coefficiente di determinazione", "r quadro" oppure "varianza comune"
- La parte complementare è chiamata "Coefficiente di indeterminazione" o "di alienazione" $(1 - r^2)$



$r^2 = 0$





Proporzione di varianza comune a due variabili

L'R^2 ci dice quanta parte della varianza di Y è "spiegata" da X.

Errore standard delle stime

- Varianza degli errori previsti

$$\frac{\sum(Y - \hat{Y})^2}{N - 2}$$

- e la relativa deviazione standard

$$s_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N - 2}} = s_y\sqrt{1 - r^2}$$

- Ne consegue che se $r = 1$, va a 0 (nessun errore)
- se $r = 0$, va a s_y (massimo errore)

Errore standard delle stime

A cosa serve l'errore standard delle stime?

- Essendo una deviazione standard, e presumendo che X e Y siano distribuite normalmente, possiamo stimare che il 95% dei valori Y stimati a partire da un certo valore X sarà compreso fra:

$$\hat{Y} - 1.96 s_{y.x} \quad \text{e} \quad \hat{Y} + 1.96 s_{y.x}$$

- dove 1.96 è il punto $|z|$ corrispondente all'area 95% attorno alla media

Test di significatività

Sui parametri della regressione semplice vengono calcolati dei test di significatività.

Un test globale (l'intero modello)

- Anova globale (quasi sempre significativa)

Un test per ogni **parametro** (intercetta, pendenza)

- viene calcolato un t-test
- se il t è significativo il parametro è statisticamente necessario per spiegare la dipendente

Viene calcolata anche la percentuale spiegata. Maggiore la %, migliore è il modello (ovvero l'ipotesi che X spieghi Y).

Sommario locale

1. Regressione lineare semplice
   - Differenza fra correlazione e influenza
   - Cos'è la regressione semplice
   - Stime, residui, pendenze
   - Test di significatività
2. In SPSS
3. Regressione semplice matriciale (M)
4. Riassunto terminologico

In SPSS: finestra principale

- `Analizza | Regressione | Lineare...`
- Trasferite in `Dipendente` la variabile che volete spiegare (la Y)
- Trasferite in `Indipendenti` la variabile che volete usare per spiegare (la X)
- Date l'`OK`

In SPSS: un esempio

Nell'esempio, Y è `Fundament`, X è `Politica`

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	78,531	1,176		66,785	,000
	politica	1,801	,210	,360	8,591	,000

a. Variabile dipendente: Fundament

- la pendenza e l'intercetta (chiamata *(Costante)*) sono presentati come non standardizzati ("B") e standardizzati ("Beta")
- Fundament = 78.531 + 1.801 * politica (non standardizzata)
- z(Fundament) = .360 * z(politica) (standardizzata)
- .360 è anche la correlazione fra Fundament e Politica

In SPSS: test

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	78,531	1,176		66,785	,000
	politica	1,801	,210	,360	8,591	,000

a. Variabile dipendente: Fundament

- test statistici: il parametro non standardizzato diviso l'errore standard produce t (1.801/0.210=8.591)
- sia la `costante` sia `Politica` sono significativamente importanti per spiegare Y tramite X

In SPSS: varianza spiegata e residui

- Un'altro risultato riportato è l'R^2 (R-quadrato) che è il quadrato della correlazione
- L'R^2 è considerato anche un indice di adeguamento (o di *fitting*) perché indica quanto "bene" funziona il nostro modello (13%)
- L'R^2 corretto considera quante indipendenti vengono utilizzate (12..8%)
- e le statistiche sui residui

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,360[a]	,130	,128	11,35156

a. Stimatori: (Costante), politica

b. Variabile dipendente: Fundament

Statistiche dei residui[a]

	Minimo	Massimo	Media	Deviazione std.	N
Valore atteso	80,332	96,5363	87,639	4,37420	498
Residuo	-38,54	57,8653	,00000	11,34013	498
Valore atteso std.	-1,670	2,034	,000	1,000	498
Residuo std.	-3,395	5,098	,000	,999	498

a. Variabile dipendente: Fundament

Sommario locale

1. Regressione lineare semplice
   - Differenza fra correlazione e influenza
   - Cos'è la regressione semplice
   - Stime, residui, pendenze
   - Test di significatività
2. In SPSS
3. Regressione semplice matriciale (M)
4. Riassunto terminologico

Immaginiamo di avere due variabili con 5 osservazioni ciascuna (prime 2 colonne) e chiamiamo l'intercetta b_0 e la pendenza b_1, per cui l'equazione $Y = a + bX + e$ diventa: $Y = b_0 + b_1 X + e$ e applichiamola ad ogni valore di X e di Y

Y	X	X^2	XY	$Y = b_0 + b_1 X + e$	
3	2	4	6	$3 = b_0 + b_1 2 + e_1$	$3 = b_0 1 + b_1 2 + e_1$
2	3	9	6	$2 = b_0 + b_1 3 + e_2$	$2 = b_0 1 + b_1 3 + e_2$
4	5	25	20	$4 = b_0 + b_1 5 + e_3$	$4 = b_0 1 + b_1 5 + e_3$
5	7	49	35	$5 = b_0 + b_1 7 + e_4$	$5 = b_0 1 + b_1 7 + e_4$
8	8	64	64	$8 = b_0 + b_1 8 + e_5$	$8 = b_0 1 + b_1 8 + e_5$
22	25	151	131	$\leftarrow$ somma	
4.4	5			$\leftarrow$ media	

Dall'espressione algebrica, passiamo ad un'espressione matriciale, perché $b_0 1 + b_1 x_i$ può essere considerato una combinazione lineare della matrice $\mathbf{X}$ dei dati per il vettore $\mathbf{b}$ dei pesi

$Y = b_0 1 + b_1 X + e$
$3 = b_0 1 + b_1 2 + e_1$
$2 = b_0 1 + b_1 3 + e_2$
$4 = b_0 1 + b_1 5 + e_3$
$5 = b_0 1 + b_1 7 + e_4$
$8 = b_0 1 + b_1 8 + e_5$

$$\begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$$

Poiché la nostra incognita è il vettore $\mathbf{b}$, dobbiamo ri-esprimerla in modo da ottenere le stime di $\mathbf{b}$

$$\widehat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix}$$

Se ricordiamo i prodotti scalari, $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}'\mathbf{y}$ vengono espressi come somme, somme al quadrato e coprodotti.

$$\begin{bmatrix} 5 & 25 \\ 25 & 151 \end{bmatrix}^{-1} \begin{bmatrix} 22 \\ 131 \end{bmatrix} \qquad \begin{bmatrix} N & \sum X \\ \sum X & \sum X^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 151/130 & -25/130 \\ -25/130 & 5/130 \end{bmatrix}$$

$$\begin{bmatrix} 151/130 & -25/130 \\ -25/130 & 5/130 \end{bmatrix} \begin{bmatrix} 22 \\ 131 \end{bmatrix} = \begin{bmatrix} \dfrac{151 * 22 - 25 * 131}{130} \\ \dfrac{-25 * 22 + 5 * 131}{130} \end{bmatrix} = \begin{bmatrix} 0.362 \\ 0.808 \end{bmatrix}$$

$$\begin{bmatrix} 0.362 \\ 0.808 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \text{intercetta} \\ \text{pendenza} \end{bmatrix}$$

$$\widehat{Y}_i = 0.362 + 0.808 X_i$$

Sommario locale

1. Regressione lineare semplice
   - Differenza fra correlazione e influenza
   - Cos'è la regressione semplice
   - Stime, residui, pendenze
   - Test di significatività
2. In SPSS
3. Regressione semplice matriciale (M)
4. Riassunto terminologico

Riassunto terminologico

- Regressione lineare semplice = regressione bivariata = predizione bivariata
- X = variabile indipendente, v. predittiva
- Y = variabile dipendente, v. predetta, v. criterio
- Y', $\hat{Y}$ = valore stimato, v. previsto
- a, b_0 = intercetta, costante
- b, b_1 = coeff. angolare, coeff. di regressione, pendenza, parametro di regressione
- β, b^* = coeff. angolare standardizzato, coeff. standardizzato